Non-linear empirical modelling to estimate phosphorus exports using continuous records of turbidity and discharge

Camille Minaudo[1], Remi Dupas[2,3], Chantal Gascuel-Odoux[2], Ophelie Fovet[2], Per-Erik Mellander[4], Philip Jordan[5], Mairead Shore[4], Florentina Moatar[1,6]

1. University François Rabelais de Tours, E.A. 6293 GéoHydrosystèmes COntinentaux, Tours, France

2. INRA, UMR1069 SAS, Agrocampus Ouest, Rennes, France

3. Department Aquatic Ecosystem Analysis and Management (ASAM), Helmholtz Centre for Environmental Research (UFZ), Magdeburg, Germany

4. Agricultural Catchments Programme, Teagasc, Environment Research Centre, Johnstown Castle, Wexford, Co. Wexford, Ireland

5. School of Geography and Environmental Sciences, Ulster University, Coleraine, N. Ireland, United Kingdom

6. OSUR – CNRS, University of Rennes1, Rennes, France

Corresponding author: camille.minaudo@univ-tours.fr

Key points

- A non-linear empirical modelling approach is presented using continuous turbidity and discharge as proxies for total and reactive P concentrations
- The best relationships between P and discharge or turbidity are non-linear with asymmetrical hysteresis
- Reconstruction of P concentration during storm events based on empirical non-linear models improves P annual load assessments

Abstract

We tested an empirical modelling approach using relatively low-cost continuous records of turbidity and discharge as proxies to estimate phosphorus (P) concentrations at a sub-hourly time step for estimating loads. The method takes into account non-linearity and hysteresis effects during storm events, and hydrological conditions variability. High-frequency records of total P and reactive P originating from four contrasting European agricultural catchments in terms of P loads were used to test the method. The models were calibrated on weekly grab sampling data combined with 10 storms surveyed sub-hourly per year (*weekly*+ survey) and then used to reconstruct P concentrations during all storm events for computing annual loads. For total P, results showed that this modelling approach allowed the estimation of annual loads with limited uncertainties (≈ -10% ± 15%), more reliable than estimations based on simple linear regressions using turbidity, based on interpolated *weekly*+ data without storm event reconstruction, or on discharge weighted calculations from weekly series or monthly series. For reactive P, load uncertainties based on the non-linear model were similar to uncertainties based on storm event reconstruction using simple linear regression (≈ 20% ± 30%), and remained lower than uncertainties obtained without storm reconstruction on weekly or monthly series, but larger than uncertainties based on interpolated *weekly*+ data (≈ -15% ± 20%). These empirical models showed we could estimate reliable P exports from non-continuous P time series when using continuous proxies, and this could potentially be very useful for completing time-series datasets in high-frequency surveys, even over extended periods.

Index terms

0402 Agricultural systems

0430 Computational methods and data processing

0470 Nutrients and nutrient cycling (4845, 4850)

1873 Uncertainty assessment (1990, 3275)

1895 Instruments and techniques: monitoring

Keywords

phosphorus, non-linear empirical modelling; high-frequency monitoring; proxies; hysteresis

1. Introduction

Phosphorus (P) concentrations in streams and rivers present a high temporal variability that can only be captured through sub-daily or even sub-hourly sampling [*Cassidy and Jordan*, 2011]. For example, P concentrations can vary by several orders of magnitude within a few hours during storm events in small rural and flashy catchments. These dynamics of P concentrations question the relevance of the monitoring strategies adopted by water authorities, for example within the EU Water Framework Directive, where P is surveyed at best on a monthly basis [*Halliday et al.*, 2015; *Skeffington et al.*, 2015]. Many authors have shown that a higher frequency monitoring would be required to: i) improve knowledge of hydrological and biogeochemical processes such as understanding P sources, mobilization and delivery processes from soils to rivers [*Halliday et al.*, 2014; *Bowes et al.*, 2015; *Dupas et al.*, 2015b, 2015c, 2017; *Mellander et al.*, 2015; *Van Der Grift et al.*, 2016] ; ii) assess stream chemical dynamics and estimate reliable chemical fluxes with limited uncertainties to evaluate the ecological status of streams [*Johnes*, 2007; *Rozemeijer et al.*, 2010; *Cassidy and Jordan*, 2011; *Jones et al.*, 2012; *Wade et al.*, 2012; *Blaen et al.*, 2016; *Rode et al.*, 2016]; iii) monitor the evolution of water quality in large rivers impacted by multiple anthropogenic activities [*Moatar et al.*, 2013; *Minaudo et al.*, 2015; *Vilmin et al.*, 2016] and their response to mitigation measures [*van Geer et al.*, 2016]. In recent years, high frequency water quality monitoring programs have been developed [*Rode et al.*, 2016], but such efforts are costly and require heavy logistics that are currently unsuitable for river basin authorities to implement.

A commonly used monitoring strategy to understand P dynamics across time scales (storm event, seasonal, inter-annual variability) is to complement regular low frequency grab sampling, typically weekly to monthly, with high-frequency sampling during selected storm events [*Ide et al.*, 2012; *Audet et al.*, 2014; *Dupas et al.*, 2015c]. Although this strategy has proved useful to understand the hydrological and biogeochemical controls on P transfer, the time series produced remain non-continuous and estimated annual P exports are associated with high uncertainties [*Defew et al.*, 2013]. Consequently, there is a need to develop appropriate methods that help to reconstruct P series during periods when no high frequency data are available, during base flow periods and unmonitored runoff events. The information contained within continuous records of parameters such as turbidity and discharge are rarely considered despite these measurements being commonly available, robust and low-cost.

A previous study has used turbidity as an explanatory variable to estimate total P concentrations with linear mixed models [*Jones et al.*, 2011]. However, this method does not account for the commonly observed hysteresis loops between P concentrations and turbidity or discharge [*Bieroza and Heathwaite*, 2015; *Bowes et al.*, 2015; *Dupas et al.*, 2015c; *Perks et al.*, 2015]. Additionally, this approach has not been tested to provide proxies of reactive phosphorus (RP) concentrations and fluxes. More recently, *Mather and Johnson* [2015] developed a non-linear empirical model to predict suspended sediment (SS) time series based on continuous discharge time series. This approach requires a limited number of continuous observation data of both the explanatory variable and the target variable, here SS, during different flow conditions to build an empirical model to estimate SS concentrations during unmonitored storm events.

In the present study, we propose to transpose this approach to P. We hypothesized that combining continuous records of turbidity and discharge with non-continuous series of P concentration (total and reactive P), with a limited number of storm events monitored at high-frequency during different hydrological conditions, could be used to calibrate non-linear empirical models and reconstruct continuous P series. The objectives were to determine i) whether this type of approach is suitable for total and/or reactive P in streams of small agricultural catchments, and ii) how many storms need to be monitored at a higher resolution (hourly) to reliably calibrate empirical non-linear models and satisfactorily predict P exports compared to the usual monthly or weekly sampling, with or without storm event monitoring. This study was undertaken using high frequency total P (TP) and reactive P (RP) time series measured in four contrasting agricultural catchments on the Atlantic seaboard of Europe (France and Ireland).

2. **Methods**

2.1. Study sites

The study used TP and RP concentrations measured in four streams at the outlet of small intensively farmed catchments on the Atlantic seaboard of Europe, two in western France (Kervidy-Naizin and Moulinet) and two in southern Ireland (Timoleague and Ballycanew).

The catchments share several physical characteristics (Table 1): they are second or third Strahler order systems, present gentle topography and are exposed to a temperate oceanic climate [*Dupas et al.*, 2015c; *Mellander et al.*, 2015, 2016]. Catchment sizes vary from 5 to 12 km² and average rainfall ranges from 862 to 1060 mm year^{-1}.

Differences exist among the study catchments with respect to land use and soil types. Three catchments with intensive dairy farming are dominated by grasslands, covering 77, 77 and 60 % of the total surface area for Timoleague, Ballycanew and Moulinet, respectively. One catchment, Kervidy-Naizin, is dominated by arable land (85% of agricultural land consists of arable crops (mainly cereals and maize) and 15% is grassland) and intensive indoor animal production (dairy cows, pigs and poultry). In Kervidy-Naizin, Moulinet and Timoleague, soils are well drained [*Molenat et al.*, 2008; *Dupas et al.*, 2017]. This contrasts with Ballycanew where 74% soils are classified as poorly drained Gley soils [*Mellander et al.*, 2016].

The hydrological variability largely differed for these catchments: in 2% of the time, 8% of the total discharge occurred in Moulinet, 10% in Timoleague, 17% in Kervidy-Naizin and 26% in Ballycanew (indicator W2, following *Moatar et al.*, [2013]). In Kervidy-Naizin, the stream is usually dry from August to October while the three other catchmentstreams are perennial.

2.2. Stream monitoring

All four catchments were equipped with an automatic gauging station (time step varying from 1 min (Kervidy-Naizin and Moulinet) to 10 min (Timoleague and Ballycanew)) for determining the discharge and with an in-situ turbidity probe (time step between 10 and 15 min). In the French catchments, the turbidity probes (PONSEL TU-NA in Kervidy-Naizin and Hydrolab HL4 in Moulinet) were situated directly in the stream water column while in the Irish catchments the probes (Hach Solitax) were located in a tank continuously filled with water pumped from the stream. Potential differences in *in-situ* and *ex-situ* installations were studied and found to give comparable results [*Sherriff et al.*, 2015]. Sub-hourly datasets were aggregated and transformed into hourly time series. Rainfall was recorded hourly in the French catchments and every 10 minutes in the Irish catchments.

The P monitoring strategies differed between the French and the Irish catchments. The French monitoring was composed of a regular survey (weekly to daily grab sampling) combined with sub-hourly sampling using ISCO 612 Full-Size Portable autosamplers during a limited number of hydrological events (approximately 10 events per year). In the Moulinet catchment, P was surveyed on a weekly basis during the period October 2007 – July 2015 and 79 storms were surveyed sub-hourly. At Kervidy-Naizin, P was surveyed on a weekly basis during the period October 2007 – October 2013, and then daily from November 2013 to July 2015. Additionally, 61 storm events were surveyed sub-hourly during the period October 2007-July 2015. For each sample, one aliquot was filtered directly on-site for soluble reactive phosphorus (SRP) analysis

(0.45 μm cellulose acetate filter), and another aliquot kept unfiltered for TP determination. Both samples were then stored at 4°C until analysis within a fortnight. Soluble reactive P was determined using colorimetry by reaction with ammonium molybdate on filtered samples (ISO 15681). Precision of SRP measurement was ±4 μg L^{-1}. TP was determined with the same method, after digestion of the unfiltered samples with potassium peroxydisulfate.

In both Irish catchments, TP and total reactive P (TRP) concentrations were recorded sub-hourly, using continuous bank-side analyzers (Hach Phosphax-Sigma instruments [*Jordan et al.*, 2007]) and then aggregated to hourly data. The data recorded during the hydrological year 2011-2012 were chosen within the present study as this period had frequent storms in both winter and summer time. The two Irish catchments have different flow controls (soil drainage) and hydrological "flashiness" and respond differently to storm events. We could, therefore, test the non-linear modelling approach for a particular challenging year in catchments of contrasting hydrology. It was assumed that TRP was approximately equivalent to SRP since it was reported in a previous study that the discharge-weighted mean SRP accounted for 98-99% of the discharge-weighted mean TRP in the Ballycanew catchment [*Shore et al.*, 2014], similar in terms of land-use to Timoleague. For consistency, RP is used here to describe this fraction in both catchments following the terminology of *Haygarth and Sharpley* [2000].

Further information on the monitoring equipment used is provided in *Dupas et al.* [2015c] for the French catchments and in *Mellander et al.* [2015, 2016] for the Irish catchments.

2.3. Storm event detection with continuous discharge records

A storm detection algorithm was developed to extract each storm event from the discharge time series. The algorithm was based on the derivative of discharge (dQ/dt) which allowed the identification of the falling and rising limbs of a given hydrological event and defined the exact start and end times of each discrete storm event (Fig. 1). When dQ/dt exceeded a calibrated threshold during a given period, it was considered to be either a rising ($dQ/dt > 2\ 10^{-3}$ mm h^{-2}) or falling limb ($dQ/dt < -1.25\ 10^{-3}$ mm h^{-2}) period. If two successive periods corresponded to a rising and falling limb, they were considered to be part of the same hydrological event, as long as the gap between these periods did not exceed 2 hours. Additionally, discharge amplitudes had to exceed 0.015 mm h^{-1} to be identified as storm events.

2.4. Non-linear empirical modelling

Several levels of analysis were conducted and presented as different layers (Fig. 2).

2.4.1. Dataset separation between calibration and evaluation datasets

The storm event datasets where split into calibration sub-datasets (Layer 1) and model evaluation sub-datasets (Layer 2).

For the French datasets, 60% of P-surveyed storms were randomly chosen among the total available data and were added to the weekly frequency monitoring; this constituted the calibration dataset. Thus, the calibration dataset at Kervidy-Naizin was composed of 37 storm events randomly selected among 61 P-surveyed events out of the 266 storm events that occurred over the entire period of record. In Moulinet, the calibration dataset was composed of 47 storm events randomly chosen among 79 P-surveyed events out of the 266 storm events that occurred over the entire period of record. The evaluation datasets were then respectively constituted by the 24 and 32 remaining storm events in Kervidy-Naizin and Moulinet.

For the Irish datasets, the continuous records of P concentrations were sub-sampled to mimic the monitoring strategy of the French catchments, i.e. a combination of a weekly sampling with a sub-hourly survey for a few storm events every year. For that purpose, a weekly survey was randomly simulated by subsampling the continuous time series every 7 days: the date of the first sample was randomly chosen among the first 7 days of the considered period, and the sampling hour was selected randomly within reasonable working hours (from 8am to 5pm). Additionally, 10 events per year were randomly chosen among the available data to compose the set of intensively surveyed events. The combination of these two samplings constituted what is hereafter called a "*weekly+*" sampling. *Weekly+* time series were then considered as calibration data and the rest of the continuous time series was the evaluation data.

Because performances by the models can be sensitive to this dataset separation step, the successive steps of data separation, calibration and evaluation were repeated 500 times. This number of successive iterations was determined based on an analysis of error distribution variations from 2 iterations to 1000 (results not shown).

2.4.2. Layer 1 – Calibration

Non-linear empirical models with hysteresis effects were developed following a similar approach to that reported by *Mather and Johnson* [2014, 2015]. These models were calibrated on each catchment dataset separately (Fig. 3).

The different models tested in this study are denoted models M1, M2 and M3 (Equations 1, 2, 3) where $P(t)$ is the P concentration (either TP or RP) at time t and $X(t)$ is the chosen explanatory variable (turbidity for TP or Q for RP) at time t, P_0 is the minimum between the observation of P before and after the P surveyed storm (i.e. baseflow concentration observed through the regular weekly sampling, or the first/last observation of the next/previous high-frequency storm event surveyed), and X_0 is the value of the chosen explanatory variable at the time corresponding to P_0.

Model M1: $$P(t) = a \cdot X(t) + b \cdot \frac{dX(t)}{dt} \qquad \text{Equation 1}$$

Model M2: $$P(t) - P_0 = a \cdot (X(t) - X_0) + b \cdot \frac{dX(t)}{dt} \qquad \text{Equation 2}$$

Model M3: $$P(t) = a \cdot X(t)^c + b \cdot \frac{dX(t)}{dt} \qquad \text{Equation 3}$$

Coefficient a describes the mean slope between $P(t)$ and $X(t)$; b describes the direction and amplitude of the hysteresis loop (clockwise if positive, counterclockwise if negative); and c describes the shape of the loop (symmetrical if equal to 1, and curved if different from 1). Model M1 predicts absolute concentrations. Model M2 is based on the hypothesis that hysteresis patterns might depend on initial turbidity or discharge conditions, or on their temporal evolution during storm events recession. Thus, M2 predicts relative variations, the baseflow value (P_0 term) being added afterwards. Model M3 considers the possibility of asymmetrical hysteresis loops. Model M1 is therefore a particular case of M3, where parameter c equals 1.

Previous studies have shown the hysteretic patterns between TP concentrations and turbidity on one side, and on RP concentrations and discharge on the other side [*Grayson et al.*, 1996; *Bowes et al.*, 2005; *Jones et al.*, 2011]. The explanatory variable X was then chosen accordingly, i.e. turbidity for TP, and discharge for RP.

Five steps were considered to apply these non-linear models (Fig. 3):

- Step 1. For each individual storm surveyed, coefficients (a, b, c) of Equations 1, 2, 3 were fitted on the calibration data series using iterative least squares estimates.
- Step 2. Because coefficients a, b, c might differ from one storm to another (e.g. due to different sources or different P transfer processes [*Bieroza and Heathwaite*, 2015]), the best calibrated sets were first selected according to a Nash-Sutcliffe criterion [*Nash and Sutcliffe*, 1970] above 0.5 and more than 5 observations within the storm event.
- Steps 3. In order to choose the right set of coefficients for a new storm event, the sets of coefficients were clustered using an agglomerative hierarchical classification, using Euclidean distance as a distance metric. The cutting threshold, i.e. the number of

clusters, was determined according to *Caliński and Harabasz* [1974] and the maximum number of clusters was set at 5. Coefficients *a*, *b*, *c* were then re-calibrated among each of the different clusters to determine a single set of coefficients representative of each cluster.

- Step 4. Decision trees were built to allocate unmonitored storm events to the previously defined clusters with given parameter values. This was based on the linkage (*Linkage* Matlab© function) between the different clusters identified previously and a set of hydrological indicators chosen to characterize the event. The hydrological indicators were the following: i) the variation of discharge during the event (Q_{max}-Q_{min}), ii) the cumulated rainfall on the day when the storm event started, iii) the cumulative rainfall over 10 days before the event, iv) the average discharge over 10 days before the event, v) the average groundwater depth in the riparian wells over 10 days before the event when data were available (i.e. at Timoleague and Kervidy-Naizin only). The first two indicators were related to the event itself, while the last three were related to antecedent catchment wetness conditions.
- Step 5. Decision trees were then used to assign a, b, c parameter values to a new storm and predict P concentrations and fluxes using the *ClassificationTree* set of functions in Matlab©. During inter-storm periods, RP and TP concentration were interpolated linearly, using observations from weekly monitoring.

2.4.3. Layer 2 - Evaluation

Performances of non-linear models were evaluated at two different time-scales (Fig. 2): i) at the storm event scale, using comparable model settings in all four catchments (same number of storms for calibration step); ii) at the annual scale in the two Irish catchments where the monitoring was near-continuous and thus allowed for calculation of actual loads on measurements.

At the storm event scale, each model was evaluated for each storm event using the calibration data series described in section 2.4.1. For each storm event, the P concentration was estimated at an hourly time step. Relative root mean square errors (%RMSE) were calculated on P loads during every storm intensively surveyed to quantify the performances of the empirical models.

The annual scale evaluation could only be conducted in the Irish catchments because of their near-continuous data. Annual loads were estimated by multiplying continuous discharge by reconstructed P concentrations estimated by models and interpolated P concentrations (after step 5, see section 2.4.2). The performances of the model at the annual time-scale were quantified using relative errors, relative bias and standard deviation of relative errors of loads.

2.4.4. Layer 3 – Comparing different strategies to assess annual loads

Performances of non-linear modelling on estimating annual loads were compared to more common ways of assessing loads, with or without storm reconstruction (Fig. 2). Again, this was conducted on the Irish dataset only (Timoleague and Ballycanew) where P measurements were near-continuous (allowing for computing the actual load). Thus, five different strategies were compared:

i) A discharge weighted load calculation based on a monthly discrete sampling. Those monthly sub-sampled time series were built following the same steps as the weekly subsampling described in section 2.4.1. Annual loads for these sub-sampled series were estimated using discharge weighted formula (Eq. 4).

$$L_y = \frac{\sum C_i Q_i}{\sum Q_i} \bar{Q} \qquad \text{Equation 4}$$

where L_y is the calculated load during year y, C_i and Q_i are the instantaneous concentration and discharge at time i and $\bar{Q}$ is the average discharge during y.

ii) A discharge weighted load calculation based on a weekly discrete sampling. Sub-sampling and load calculation methods were similar to the monthly strategy.

iii) A simple linear interpolation between observations of a *weekly+* sampling without storm-reconstruction. Corresponding loads integrated only the storm events that were sampled and neglected the others.

iv) A *weekly+* sampling with storm-reconstruction based on a linear regression model were continuous records of turbidity and discharge were used as proxies for, respectively, TP and RP, as in non-linear models M1, M2 and M3. This model did not consider hysteresis cycles. The relationship between P concentration and the explanatory variable X followed a linear relationship according to the Equation 5 formulation.

$$\text{Linear model: } P(t) = a \cdot X(t) + b \qquad \text{Equation 5}$$

Coefficients a and b in each case were fitted by minimalizing squared errors based on the entire calibration dataset. This model was a simpler version of the model presented in the *Jones et al.* [2011] study where turbidity was used as a proxy for high-frequency TP.

v) Our approach, i.e. a *weekly*+ sampling with storm-reconstruction, based on the non-linear modelling approach developed in this study (see section 2.4.2.)

The same sensitivity test as conducted for model evaluation was run by repeating 500 times the successive steps: random calibration dataset selection, model calibration, annual load estimations and performance evaluation.

2.4.5. Layer 4 – Sensitivity analysis of non-linear models

Additionally to the sensitivity of model performances to calibration datasets, we assessed the impact of the number of P surveyed storms included in the calibration dataset on annual load estimations (Fig. 2). It was chosen to estimate model performances when the calibration dataset was composed of 6 to 20 storm events per year. This allowed an estimation of the differences in the model efficiency when more information was added in the input dataset. This was conducted with the Irish catchments' data, and compared to load assessments from a simple linear regression between turbidity and TP and between discharge and RP (see sections 2.4.2 and 2.4.4 for models constructions).

2.4.6. Layer 5 – Model application to improve P exports assessment in catchments where P is non-continuously surveyed

The model providing the best performances on P load assessment was used to estimate annual TP and RP exports in the two French catchments where P surveys are non-continuous (Fig. 2). Uncertainty was associated with these estimations based on the load uncertainties computed from the analysis made on the continuous Irish datasets at the annual scale, as errors in both Irish catchments were similar.

3. Results

3.1. Contrasting P concentration in the four catchments

Phosphorus variability and composition were different in the four catchments (Table 2). TP median concentrations ranged between 0.06 and 0.20 mg P L^{-1}, the highest concentrations being observed in the Moulinet catchment (90^{th} percentile was 0.9 mg P L^{-1} against 0.16-0.37 mg P

L^{-1} in the other catchments). RP median concentrations ranged between 0.01 and 0.05 mg P L^{-1}, the highest concentrations being comparable in Timoleague, Ballycanew and Kervidy-Naizin (0.09-0.11 mg P L^{-1}) and much lower in Moulinet (0.04 mg P L^{-1}). The proportion of RP in TP also differed in the four catchments. For example, during storm events, the RP fraction of the TP concentration represented on average approximately 40% in Timoleague, Ballycanew and Kervidy-Naizin, and sometimes up to 80% in Kervidy-Naizin. In Moulinet, RP represented less than 10% of TP most of the time, especially during storm events, and concentrations remained under 0.06 mg RP L^{-1}. Ninety percent of the annual TP load occurred in 51% of the time in Timoleague against 21% in Ballycanew. For annual RP loads, this was 54% of the time in Timoleague against 34% in Ballycanew.

3.2. Storm events characteristics in the four catchments

The algorithm identified 266 and 329 storm events in Kervidy-Naizin and Moulinet, respectively, over the entire period, i.e. approximately 38 and 47 storms per year respectively (Table 2). In the Irish catchment during the 2011-2012 hydrological year, the algorithm identified 38 and 49 storms in Timoleague and Ballycanew, respectively. Storm event amplitudes were larger in Ballycanew than in the other catchments: among all the events identified, 12% of events exhibited specific discharge amplitudes over 0.1 mm h^{-1} at Moulinet, against 29% at Kervidy-Naizin, 39% at Timoleague and only 49% at Ballycanew. Storm events were longer in Timoleague and Ballycanew than in Kervidy-Naizin and Moulinet: event durations ranged between a few hours and several days. Average event duration was 18 hours at Moulinet, 30 hours at Kervidy-Naizin, and 42 hours at Timoleague and Ballycanew. Approximately 95% events lasted less than 3 days in the different catchments, except at Timoleague where the proportion was 87%.

3.3. Empirical models performances during calibration step

The three different mathematical formulations used to calibrate non-linear models using turbidity as a proxy for TP and discharge as a proxy for RP were tested on all available intensively surveyed storms. The distribution of Nash-Sutcliffe (NS) criterions computed for each storm individually were very low for the symmetrical hysteresis models M1 and M2, and were for most of the time below 0.5 independent of catchment or variable (TP or RP) (Figure 4). Only a small percentage of storms could be considered for further model calibration steps, indicating that non-linear models considering symmetrical hysteresis poorly fitted the observations. The asymmetrical hysteresis model M3, however, provided NS values most of the time over 0.5, and a large percentage of storms could be used for the next calibration steps.

Thus, the rest of the study focused on both TP and RP in all 4 catchments based on the non-linear model with asymmetrical hysteresis loops (M3). Models M1 and M2 are no longer used or reported hereafter.

3.4. Performances on predicting P concentration and fluxes at different time scales

3.4.1. Performances at the storm event scale

Errors at the storm event scale for predicting TP and RP fluxes from model M3 were large (Table 3). For TP, medians over 500 iterations of relative RMSE (%RMSE) ranging between 51 and 104%. Variability through the different simulations were considerable. The number of simulations providing %RMSE for TP flux at the storm event scale under 50% was small with, respectively, 49, 2, 11 and 9% for Timoleague, Ballycanew, Kervidy-Naizin and Moulinet. Most simulations provided %RMSE for TP fluxes under 100% in the Irish catchments, but error ranges were higher in the French catchments with 90th percentile on %RMSE reaching 129% in Kervidy-Naizin and up to 193% in Moulinet. Similar values were found for RP fluxes. The non-linear modelling approach showed unacceptable %RMSE values for predicting RP loads in Moulinet catchment (median %RMSE was 238%), but median %RMSE in the other catchments ranged between 72 and 79%. The number of simulations providing %RMSE for RP flux at the storm event scale under 50% was, respectively, 12, 26, 5 and 0% for Timoleague, Ballycanew, Kervidy-Naizin and Moulinet.

Continuous series reconstructed by the non-linear model M3 preserved storm event concentrations dynamics (Fig. 5). If peak amplitudes were subject to large errors, especially for RP, peak phases corresponded to the observed concentrations. Predictions over 500 iterations were variable, and uncertainties depended on the storm event considered.

3.4.2. Performances at the annual scale

For model evaluation, annual load estimations could be calculated for the Irish catchments only. Errors were relatively low (Fig. 6). For annual TP load prediction, 10th to 90th percentile range of relative error was -5 to +18% for Timoleague and -26 to +1% for Ballycanew. This corresponded to relative bias ± s.d. error of 7% ± 12% in Timoleague and -11% ± 17% in Ballycanew. In Timoleague, we counted in results from the non-linear modelling that 60%

simulations out of 500 iterations produced relative errors on TP annual loads included within the range ± 10%. The proportion was 35% in Ballycanew.

For RP, non-linear model M3 tended to overestimate the annual load: 10^{th}-90^{th} percentile errors ranged between -5 to +48% (bias ± imprecision were approximatively 20% ± 30%). In Timoleague, we counted that 42% simulations out of 500 iterations produced relative errors on RP annual loads included within the range ± 10%. The proportion was 38% in Ballycanew.

3.5. Comparison of five different strategies to estimate annual loads

3.5.1. Comparison with linear regression models

Simple linear regression models using continuous records of turbidity and discharge respectively as proxies for TP and RP exhibited variable coefficients of determination (results shown in a Supplement information S1): R^2 between turbidity and TP concentration extracted from the calibration dataset ranged throughout the 500 iterations between 0.5 and 0.8 in Timoleague and between 0.2 and 0.7 in Ballycanew; R^2 between discharge and RP concentration ranged between 0 and 0.65 in Timoleague and between 0.15 and 0.6 in Ballycanew.

When used to reconstruct TP and RP concentrations during storm events and estimate annual loads, these simple regressions provided load estimates associated with larger uncertainties than with the non-linear modelling approach. The simple linear method tended to underestimate TP (bias ± imprecision was approximatively 15% ± 20% at both sites) and overestimate RP (bias ± imprecision was 29% ± 35% in Timoleague and 16% ± 24% in Ballycanew). A smaller number of simulations provided annual load estimates within the range ± 10%: in Timoleague, 41% of simulations were within this range for TP (against 60% with the non-linear model M3) and 19% for RP (against 42% with M3); in Ballycanew, it was 42% for TP (against 42% with M3), and 30% for RP (against 38% with M3). At the scale of the storm event, it appeared that, even if the two or three most contributing events were better predicted with the simple linear model, most event fluxes were more reliably predicted with the non-linear model (results can be found in Supporting information S2).

3.5.2. Comparison with simple interpolation of measurements from different sampling strategies

Using simple linear interpolation of measurement without reconstruction of storm event concentrations, the *weekly+,* weekly, and monthly strategies were subject to large errors and tended to underestimate annual loads: for both TP and RP, 10^{th}-90^{th} percentile errors ranged between -40 to -1% for a *weekly+* strategy, -40 to +40% for a weekly sampling, and -50 to +35% for a monthly survey. Bias ranged between -34 to -7%, and the smallest bias was obtained with a weekly sampling strategy, but was associated with a 38% imprecision. Standard deviation errors ranged between 16 and 55%: the highest values resulted from the lowest sampling frequencies.

3.6. Sensitivity of the empirical models to the calibration dataset

Results have shown how much the performance of empirical modelling of TP using turbidity and RP using discharge largely differed depending on the 500 random draws that were made to separate calibration and evaluation datasets. Models were sensitive to the information contained initially in the calibration dataset, but all these results originated from the hypothesis that 10 storms intensively surveyed per year should be enough. To assess the sensitivity of non-linear modelling to the quantity of information contained into calibration data, an analysis was conducted on the number of storms initially included in the calibration dataset. This was tested at the annual scale, based on the continuous records available in the Irish catchments.

The number of events contained initially in the calibration dataset highly changed the quality of annual load predictions (Fig. 7). Both bias and imprecision were reduced when using a larger calibration dataset. In Timoleague, errors on annual load estimations of TP using the non-linear model decreased from -1% ± 18% to less than 5% ± 8% when using 6 to 20 storms among 38. Predictions also improved for RP loads estimations in Timoleague: errors reduced from 51% ± 99% to 11% ± 32%. In Ballycanew, TP errors reduced from -12% ± 19% to 8% ± 12% and RP errors reduced from 33% ± 51% to 9% ± 15%.

3.7. Using non-linear empirical modelling to improve annual load assessment in catchments where P was non-continuously surveyed

The empirical models enabled the calculation of continuous series of TP using all the information contained in the available data in the French catchments, i.e. 266 and 329 events for Kervidy-Naizin and Moulinet respectively. Based on the non-linear modelling technique developed in this study, TP annual loads ranged between 18 and 63 kg P $year^{-1}$ km^{-2} in Kervidy-Naizin and between 30 and 65 kg P $year^{-1}$ km^{-2} in Moulinet, depending on the year (Fig. 8). The proportion of RP in the total annual load based on the model ranged between 13 and 48 % in Kervidy-Naizin depending on the year, and remained under 5% in Moulinet. Although P exports were quite similar between the two catchments, a larger part of the annual TP load occurred in Kervidy-Naizin during storm events: on average 62% versus 51% in Moulinet. In Kervidy-Naizin, 73% of the RP annual load was exported during storms. In Moulinet, 19% of the small amount of RP load was exported during storm events.

Compared to load estimations with storm event reconstructions, the *weekly*+ strategy globally underestimated TP load values, with a much larger uncertainty window. Differences between loads assessed with the *weekly*+ survey, or assessed based on the non-linear empirical model, were even larger in Moulinet: TP loads with the non-linear model were three to seven-fold of the estimated load without storm reconstruction for the years 2012 and 2014.

4. Discussion

4.1. Should we use turbidity and discharge as proxies for TP and RP?

This study showed that storm event reconstruction based on the association of proxies (continuous turbidity for TP), a *weekly*+ survey (i.e. a weekly sampling added to 10 storms intensively surveyed per year), and non-linear empirical modelling provided more reliable annual load predictions for TP compared to simple discharge weighted load calculations or compared to continuous series based on linear regressions between turbidity and TP.

For RP, our empirical modelling approach based on 10 storms per year and continuous discharge used as proxy did not improve load assessments since predictions at the storm event scale were subject to large errors and provoked at least 15% ± 25% errors on annual loads. In the case of RP, simple calculations based on *weekly*+ datasets remained the best choice. These results show a lower predictability of RP by the hydrological proxy we used, probably due to direct effects of human activities occurring mainly in spring (e.g. manure spreading, mineralization of organic matter), as indicated by *Dupas et al.* [2016a].

However, load estimations were highly dependent on the set of storm events used for calibrating the non-linear model: even for RP, some predictions could be very good as we counted in both Irish catchments that around 40% of simulations (among 500 iterations) produced errors included within the reasonable range ±10%. Therefore, further analysis should be done to determine which set of storms has to be selected to produce the lowest load errors. Additionally, results showed that when the number of storms included in the calibration of the non-linear model was increased, errors were highly reduced for both TP and RP load estimations. One can expect in non-continuous P series recorded over several years with 10 storm events intensively surveyed per year would allow non-linear empirical models to provide more reliable annual loads.

Empirical models are useful tools to assess P exports in small agricultural catchments. This study strongly recommends stakeholders to develop monitoring strategies that combine weekly and a selection of sub-hourly storm samplings (*weekly+*). This will considerably help to assess P exports from, at least, small agricultural catchments where diffuse exports associated with storm events is dominant. This type of monitoring appears costly but provides useful information to improve understanding of catchment behavior and P export assessment: in the empirical approach developed here TP loads are reasonably well estimated, even in catchments with proportionally large RP concentrations that are more difficult to estimate.

Based on this study, catchment managers would then have to deploy a *weekly+* strategy with approximately 10 storms intensively surveyed per year over at least two years to cover the diversity of hydrological and agricultural conditions, depending on the inter-annual climate variability. Then, TP load estimations would be predicted for the first two years and the subsequent years with limited uncertainties (≈ -10 ± 10%) using non-linear modelling applied on continuous turbidity data, which is likely to be cheaper and straightforward compared to high-frequency P surveys over the entire extended period. Because P concentration relationship with turbidity or discharge may not be stable after implementation of mitigation measures in the catchment, additional control monitoring would then need to be set up, to control and/or recalibrate the empirical models, as it is usually conducted for discharge rating curves. This would require sampling a few storm events per year.

To limit prediction errors on load calculations, the hydrological events intensively surveyed must be targeted according to the diversity of storm event typologies existing, and ideally characterized in beforehand. Further work should be done, but it seems reasonable to assume these events have to be spread across the period of record, through different climatic and agricultural seasons but also a few events have to be consecutive in order to represent different catchment wetness conditions. Apart from a peculiar event such as an uncontrolled point-source loading, the calibration dataset must include events of different amplitudes and in different seasons, so it is likely that model predictions could cover the variability of conditions encountered in study catchments. Thus, to proceed properly, monitoring for modelling programs would require (i) hydro-meteorological records to be able to characterize the variability of storm events within a year and inter-annually; and (ii) hydrochemical records to be representative of this variability, associated with continuous records of a relevant proxy (turbidity). Achieving this, the use of empirical models can be a relevant compromise for estimating annual P loads, providing more reliable estimates than calculations based on a low frequency sampling and more affordable than direct continuous monitoring of P concentrations.

4.2. New insights about P export regime in catchments where P is non-continuously surveyed

Continuous series of TP and RP were reconstructed for non-continuous P series (in the two French catchments) based on the non-linear empirical models and all data available. These synthetic series provided new knowledge on mean level and inter-annual variability of P exports in these catchments. Results in the present study show that P export estimations without storm event reconstruction lead to large errors, and estimations based on empirical modelling are more reliable. It was estimated with the non-linear model in Kervidy-Naizin that, depending on the considered year, 13 to 49% of TP load was composed by RP fraction, 24% on average over the study period. The highest proportion (49%) was calculated for a particularly wet year in Kervidy-Naizin (1219 mm in 2013 versus 924 mm on average), suggesting more RP transport probably due to soil-groundwater interactions taking place during longer periods and over large areas, previously identified as the mechanism controlling soluble P transport, [*Dupas et al.*, 2015a, 2015b, 2017]. The annual TP exports from Moulinet was similar to that in Kervidy-Naizin, but the proportion of RP was smaller (on average, 9%) . RP concentrations are subjected to high errors due to analytical techniques and storage [*Jarvie et al.*, 2002]; thus, the main limitation for estimating annual RP loads in this catchment might be linked to measurement uncertainties [*Dupas et al.*, 2016b]. Improving data quality is crucial before being able to calibrate a reliable model. In this way, bankside analyzers constitute a good solution, especially because P analysis would be immediate (no sample decay during storage), and filtration would not be delayed, limiting the risk of adsorption to particles when samples stay several days in autosampler bottles [*Jordan et al.*, 2007].

Strong disparities could be found between the two catchments considering the very different proportion of P load occurring during storm events only, since it was found that 50 to 90% of the P exports occurred during storm events in Kervidy-Naizin, contrasting with Moulinet where it was 30 to 60%. This is concomitant with the observation made on discharge variability: discharge in Moulinet presented the lowest hydrological reactivity index W2 (8%, Table 1), and despite most P exports were transferred as particulate P, fluxes during low flows should not be ignored.

4.3. Potential improvements in the empirical approach

It is clear that empirical models strongly depend on the calibration step. Selecting the set of storms intensively surveyed and used for model calibration appears crucial. This is likely to be the key to improve this approach, and further analysis should try to answer the two following questions: based on hydrological indicators, what constitutes the best set of surveyed storms to minimize load prediction errors? And, can we predict confidently that these optimal hydrological conditions will occur and choose whether or not autosamplers have to be triggered for the next storm event?

Other explanatory variables than turbidity and discharge could have been tested to predict RP concentrations and fluxes. For example, continuous measurements of electrical conductivity or spectrometer data can also provide good results for RP as shown by *Etheridge et al.* [2014]. A combination of several parameters could also be used as explanatory variables, to provide as much information as possible to the models. Additionally, other mathematical equations have been proposed to represent the hysteresis effects between two variables. For example, *Mather and Johnson*, [2014] proposed a more complex equation than model M3 (Eq. 3) to predict suspended solids concentration based on turbidity in which several terms help to describe as best as possible non-linearity and complex hysteresis loops.

Alternative methods such as Partial Least Squares models [*Wold et al.*, 2001] or machine learning methods might provide good performances on predicting P concentrations and loads. This has already been developed for predicting suspended sediment concentrations and fluxes [*Onderka et al.*, 2012; *Ouellet-Proulx et al.*, 2016] but hasn't been tested yet to assess P exports. Since we show that the models' performances are site-dependent, the different existing methods (including the empirical models tested within our study) would have to be tested specifically on each catchment.

5. Conclusion

The non-linear empirical modelling approach developed in this study showed that the use of continuous low-cost measurements such as turbidity and discharge can be useful to help predict reliable estimates of P exports. For predicting TP loads empirical models applied on weekly data combined with 10 storms intensively surveyed per year (*weekly*+ survey) allowed the estimation of annual loads with limited uncertainties (≈ 10 ± 15% errors), more reliable than estimations based on monthly series (≈ -30 ± 50%), weekly series (≈ -10 ± 35%), or based on the *weekly*+ data without storm event reconstruction (≈ -25 ± 30%) or with simple regression models using turbidity and discharge to reconstruct P variations during storm events (≈ 15 ±

20%). For reactive P, load uncertainties based on non-linear empirical models were larger than uncertainties based on *weekly+* data without storm reconstructions (≈ 20 ± 30%), although, it was shown that empirical models statistically provide the best results.

This study showed that the asymmetrical non-linear model (M3) provided the best representation of TP-turbidity and RP-discharge hysteresis cycles and was convenient for most sites. The method developed here would largely benefit being tested on other sites with high-frequency datasets and contrasting catchments.

Acknowledgement

The Matlab code developed in this study is available in the Supporting information or can be requested from the corresponding author. This work was funded by the “Agence de l’Eau Loire Bretagne” (Loire and Brittany water basin authority) via the TRANS-P project. Long-term monitoring in the Kervidy-Naizin and Moulinet catchments were supported by ORE AgrHyS and ORE PFC. French datasets are available at http://www6.inra.fr/ore_agrhys/ or upon request to agrhys@inra.fr. The Irish monitoring is part of the Agricultural Catchments Programme (ACP) but does not share data publicly.

References

Audet, J., L. Martinsen, B. Hasler, H. De Jonge, E. Karydi, N. B. Ovesen, and B. Kronvang (2014), Comparison of sampling methodologies for nutrient monitoring in streams: Uncertainties, costs and implications for mitigation, *Hydrol. Earth Syst. Sci.*, *18*(11), 4721–4731, doi:10.5194/hess-18-4721-2014.

Bieroza, M. Z., and a. L. Heathwaite (2015), Seasonal variation in phosphorus concentration–discharge hysteresis inferred from high-frequency in situ monitoring, *J. Hydrol.*, *524*, 333–347, doi:10.1016/j.jhydrol.2015.02.036.

Blaen, P. J., K. Khamis, C. E. M. Lloyd, C. Bradley, D. Hannah, and S. Krause (2016), Real-time monitoring of nutrients and dissolved organic matter in rivers: Capturing event dynamics, technological opportunities and future directions, *Sci. Total Environ.*, *569–570*, 647–660, doi:10.1016/j.scitotenv.2016.06.116.

Bowes, M. J., W. A. House, R. A. Hodgkinson, and D. V. Leach (2005), Phosphorus-discharge hysteresis during storm events along a river catchment: The River Swale, UK, *Water Res.*, *39*(5), 751–762, doi:10.1016/j.watres.2004.11.027.

Bowes, M. J., H. P. Jarvie, S. J. Halliday, R. a. Skeffington, a. J. Wade, M. Loewenthal, E. Gozzard, J. R. Newman, and E. J. Palmer-Felgate (2015), Characterising phosphorus and nitrate inputs to a rural river using high-frequency concentration–flow relationships, *Sci. Total Environ.*, *511*, 608–620, doi:10.1016/j.scitotenv.2014.12.086.

Caliński, T., and J. Harabasz (1974), A dendrite method for cluster analysis, *Commun. Stat.*, *3*(1), 1–27, doi:10.1080/03610927408827101.

Cassidy, R., and P. Jordan (2011), Limitations of instantaneous water quality sampling in surface-water catchments: Comparison with near-continuous phosphorus time-series data, *J. Hydrol.*, *405*(1–2), 182–193, doi:10.1016/j.jhydrol.2011.05.020.

Defew, L. H., L. May, and K. V. Heal (2013), Uncertainties in estimated phosphorus loads as a function of different sampling frequencies and common calculation methods, *Mar. Freshw. Res.*, *64*, 373–386, doi:10.1071/MF12097.

Dupas, R., C. Gascuel-Odoux, N. Gilliet, C. Grimaldi, and G. Gruau (2015a), Distinct export dynamics for dissolved and particulate phosphorus reveal independent transport mechanisms in an arable headwater catchment, *Hydrol. Process.*, *29*(14), 3162–3178, doi:10.1002/hyp.10432.

Dupas, R., G. Gruau, S. Gu, G. Humbert, A. Jaffrézic, and C. Gascuel-Odoux (2015b), Groundwater control of biogeochemical processes causing phosphorus release from riparian wetlands, *Water Res.*, *84*(September), 307–314, doi:10.1016/j.watres.2015.07.048.

Dupas, R., R. Tavenard, O. Fovet, N. Gilliet, C. Grimaldi, and C. Gascuel-Odoux (2015c), Identifying seasonal patterns of phosphorus storm dynamics with dynamic time warping, *Water Resour. Res.*, *51*(11), 8868–8882, doi:10.1002/2015WR017338.

Dupas, R., S. Jomaa, A. Musolff, D. Borchardt, and M. Rode (2016a), Disentangling the influence of hydroclimatic patterns and agricultural management on river nitrate dynamics from sub-hourly to decadal time scales, *Sci. Total Environ.*, *571*, 791–800, doi:10.1016/j.scitotenv.2016.07.053.

Dupas, R., J. Salmon-Monviola, K. J. Beven, P. Durand, P. M. Haygarth, M. J. Hollaway, and C. Gascuel-Odoux (2016b), Uncertainty assessment of a dominant-process catchment model of dissolved phosphorus transfer, *Hydrol. Earth Syst. Sci.*, *20*(12), 4819–4835, doi:10.5194/hess-20-4819-2016.

Dupas, R., P.-E. Mellander, C. Gascuel-Odoux, O. Fovet, E. B. McAleer, N. T. McDonald, M. Shore, and P. Jordan (2017), The role of mobilisation and delivery processes on contrasting dissolved nitrogen and phosphorus exports in groundwater fed catchments, *Sci. Total Environ.*, *599*, 1275–1287, doi:10.1016/j.scitotenv.2017.05.091.

Etheridge, J. R., F. Birgand, J. a. Osborne, C. L. Osburn, M. R. Burchell Ii, and J. Irving (2014), Using in situ ultraviolet-visual spectroscopy to measure nitrogen, carbon, phosphorus, and suspended solids concentrations at a high frequency in a brackish tidal marsh, *Limnol. Oceanogr. Methods*, *12*, 10–22, doi:10.4319/lom.2014.12.10.

van Geer, F. C., B. Kronvang, and H. P. Broers (2016), High resolution monitoring of nutrients in groundwater and surface waters: process understanding, quantification of loads and concentrations and management applications, *Hydrol. Earth Syst. Sci. Discuss.*, (M), 1–21, doi:10.5194/hess-2016-169.

Grayson, R., B. Finlayson, C. Gippel, and B. Hart (1996), The potential of field turbidity measurements for the computation of total phosphorus and suspended solids loads, *J.*

Environ. Manage., *47*, 257–267, doi:10.1006/jema.1996.0051.

Van Der Grift, B., H. Peter Broers, W. Berendrecht, J. Rozemeijer, L. Osté, and J. Griffioen (2016), High-frequency monitoring reveals nutrient sources and transport processes in an agriculture-dominated lowland water system, *Hydrol. Earth Syst. Sci.*, *20*(5), 1851–1868, doi:10.5194/hess-20-1851-2016.

Halliday, S., R. Skeffington, M. Bowes, E. Gozzard, J. Newman, M. Loewenthal, E. Palmer-Felgate, H. Jarvie, and A. Wade (2014), The Water Quality of the River Enborne, UK: Observations from High-Frequency Monitoring in a Rural, Lowland River System, *Water*, *6*(1), 150–180, doi:10.3390/w6010150.

Halliday, S. J., R. a. Skeffington, A. J. Wade, M. J. Bowes, E. Gozzard, J. R. Newman, M. Loewenthal, E. J. Palmer-Felgate, and H. P. Jarvie (2015), High-frequency water quality monitoring in an urban catchment: hydrochemical dynamics, primary production and implications for the Water Framework Directive, *Hydrol. Process.*, *29*(15), 3388–3407, doi:10.1002/hyp.10453.

Haygarth, P. M., and A. N. Sharpley (2000), Terminology for Phosphorus Transfer, *J. Environ. Qual.*, *29*(1), 10–15.

Ide, J., M. Chiwa, N. Higashi, R. Maruno, Y. Mori, and K. Otsuki (2012), Determining storm sampling requirements for improving precision of annual load estimates of nutrients from a small forested watershed, *Environ. Monit. Assess.*, *184*, 4747–4762, doi:10.1007/s10661-011-2299-9.

Jarvie, H., P. Withers, and C. Neal (2002), Review of robust measurement of phosphorus in river water: sampling, storage, fractionation and sensitivity, *Hydrol. Earth Syst. Sci.*, *6*(1), 113–131, doi:10.5194/hess-6-113-2002.

Johnes, P. J. (2007), Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodology, sampling frequency, baseflow index and catchment population density, *J. Hydrol.*, *332*(1–2), 241–258, doi:10.1016/j.jhydrol.2006.07.006.

Jones, A. S., D. K. Stevens, J. S. Horsburgh, and N. O. Mesner (2011), Surrogate Measures for Providing High Frequency Estimates of Total Suspended Solids and Total Phosphorus Concentrations, *JAWRA J. Am. Water Resour. Assoc.*, *47*(2), 239–253, doi:10.1111/j.1752-1688.2010.00505.x.

Jones, A. S., J. S. Horsburgh, N. O. Mesner, R. J. Ryel, and D. K. Stevens (2012), Influence of Sampling Frequency on Estimation of Annual Total Phosphorus and Total Suspended Solids Loads, *J. Am. Water Resour. Assoc.*, *48*(6), 1258–1275, doi:10.1111/j.1752-1688.2012.00684.x.

Jordan, P., A. Arnscheidt, H. McGrogan, and S. McCormick (2007), Characterising phosphorus transfers in rural catchments using a continuous bank-side analyser, *Hydrol. Earth Syst. Sci.*, *11*(1), 372–381, doi:10.5194/hess-11-372-2007.

Mather, A. L., and R. L. Johnson (2014), Quantitative characterization of stream turbidity-discharge behavior using event loop shape modeling and power law parameter decorrelation, *Water Resour. Res.*, *50*(10), 7766–7779, doi:10.1002/2014WR015417.

Mather, A. L., and R. L. Johnson (2015), Event-based prediction of stream turbidity using regression and classification tree approaches, *J. Hydrol.*, *in review*, 751–761, doi:10.1016/j.jhydrol.2015.10.032.

Mellander, P.-E., P. Jordan, M. Shore, N. T. Mcdonald, D. P. Wall, G. Shortle, and K. Daly (2016), Identifying contrasting influences and surface water signals for specific groundwater phosphorus vulnerability, *Sci. Total Environ.*, *541*, 292–302, doi:10.1016/j.scitotenv.2015.09.082.

Mellander, P. E., P. Jordan, M. Shore, A. R. Melland, and G. Shortle (2015), Flow paths and phosphorus transfer pathways in two agricultural streams with contrasting flow controls, *Hydrol. Process.*, *3518*(January), 3504–3518, doi:10.1002/hyp.10415.

Minaudo, C., M. Meybeck, F. Moatar, N. Gassama, and F. Curie (2015), Eutrophication mitigation in rivers: 30 years of trends in spatial and seasonal patterns of biogeochemistry of the Loire River (1980–2012), *Biogeosciences*, *12*(8), 2549–2563, doi:10.5194/bg-12-2549-2015.

Moatar, F., M. Meybeck, S. Raymond, F. Birgand, and F. Curie (2013), River flux uncertainties predicted by hydrological variability and riverine material behaviour, *Hydrol. Process.*, *27*(25), 3535–3546, doi:10.1002/hyp.9464.

Molenat, J., C. Gascuel-Odoux, L. Ruiz, and G. Gruau (2008), Role of water table dynamics on stream nitrate export and concentration in agricultural headwater catchment (France), *J. Hydrol.*, *348*(3–4), 363–378, doi:10.1016/j.jhydrol.2007.10.005.

Nash, J. E., and J. V Sutcliffe (1970), River Flow Forecasting Through Conceptual Models Part I-A Discussion of Principles, *J. Hydrol.*, *10*, 282–290, doi:10.1016/0022-1694(70)90255-6.

Onderka, M., A. Krein, S. Wrede, N. Martínez-Carreras, and L. Hoffmann (2012), Dynamics of storm-driven suspended sediments in a headwater catchment described by multivariable modeling, *J. Soils Sediments*, *12*(4), 620–635, doi:10.1007/s11368-012-0480-6.

Ouellet-Proulx, S., A. St-Hilaire, S. C. Courtenay, and K. A. Haralampides (2016), Estimation of suspended sediment concentration in the Saint John River using rating curves and a machine learning approach, *Hydrol. Sci. J.*, *61*(10), 1847–1860, doi:10.1080/02626667.2015.1051982.

Perks, M. T., G. J. Owen, C. M. H. Benskin, J. Jonczyk, C. Deasy, S. Burke, S. M. Reaney, and P. M. Haygarth (2015), Dominant mechanisms for the delivery of fine sediment and phosphorus to fluvial networks draining grassland dominated headwater catchments, *Sci. Total Environ.*, *523*, 178–190, doi:10.1016/j.scitotenv.2015.03.008.

Rode, M. et al. (2016), Sensors in the stream: the high-frequency wave of the present, *Environ. Sci. Technol.*, acs.est.6b02155, doi:10.1021/acs.est.6b02155.

Rozemeijer, J. C., Y. Van Der Velde, F. C. Van Geer, G. H. De Rooij, P. J. J. F. Torfs, and H. P. Broers (2010), Improving load estimates for NO3 and P in surface waters by characterizing the concentration response to rainfall events, *Environ. Sci. Technol.*, *44*(16), 6305–6312, doi:10.1021/es101252e.

Sherriff, S. C., J. S. Rowan, A. R. Melland, P. Jordan, O. Fenton, and D. O. Huallacháin (2015), Investigating suspended sediment dynamics in contrasting agricultural catchments using ex situ turbidity-based suspended sediment monitoring, *Hydrol. Earth Syst. Sci.*, *19*(8), 3349–3363, doi:10.5194/hess-19-3349-2015.

Shore, M., P. Jordan, P. E. Mellander, M. Kelly-Quinn, D. P. Wall, P. N. C. Murphy, and A. R. Melland (2014), Evaluating the critical source area concept of phosphorus loss from soils to water-bodies in agricultural catchments, *Sci. Total Environ.*, *490*, 405–415, doi:10.1016/j.scitotenv.2014.04.122.

Skeffington, R. A., S. J. Halliday, A. J. Wade, M. J. Bowes, and M. Loewenthal (2015), Using high-frequency water quality data to assess sampling strategies for the EU Water

Framework Directive, *Hydrol. Earth Syst. Sci.*, *19*(5), 2491–2504, doi:10.5194/hess-19-2491-2015.

Vilmin, L., N. Flipo, N. Escoffier, and A. Groleau (2016), Estimation of the water quality of a large urbanized river as defined by the European WFD: what is the optimal sampling frequency ?, *Environ. Sci. Pollut. Res.*, doi:10.1007/s11356-016-7109-z.

Wade, A. J. et al. (2012), Hydrochemical processes in lowland rivers: Insights from in situ, high-resolution monitoring, *Hydrol. Earth Syst. Sci.*, *16*, 4323–4342, doi:10.5194/hess-16-4323-2012.

Wold, S., M. Sjöström, and L. Eriksson (2001), PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, *58*(2), 109–130, doi:10.1016/S0169-7439(01)00155-1.

Table 1. Study sites characteristics. S: catchment area, q: specific discharge (annual mean ± standard deviation), W2: percentage of water flux passing in 2% of the time [*Moatar et al.*, 2013].

	Timoleague (IR)	Ballycanew (IR)	Kervidy-Naizin (FR)	Moulinet (FR)
S (km^2)	8	12	5	5
q (mm)	417 ± 182	373 ± 129	316 ± 151	371 ± 77
W2 (%)	10	26	17	8
average rainfall (mm $year^{-1}$)	1047	1060	924	862
P concentration temporal resolution	hourly	hourly	weekly (2007-2013) daily (2013-2015) + 61 storms sub-hourly	weekly + 79 storms sub-hourly
data extent	Oct. 2011 - Sept. 2012	Oct. 2011 - Sept. 2012	Oct. 2007 - July 2015	Oct. 2007 - July 2015

Table 2. Characteristics of P concentration and load at the different study sites, and characteristics of storm events identified by the algorithm. $f_{L90\%}$: P load dynamic indicator such as 90% of the annual load occurs in $f_{L90\%}$ % of the time.

	Timoleague (IR)	Ballycanew (IR)	Kervidy-Naizin (FR)	Moulinet (FR)
TP concentration (mg P L^{-1}) median (10^{th}; 90^{th})	0.06 (0.05; 0.16)	0.07 (0.05; 0.20)	0.07 (0.02; 0.37)	0.20 (0.03; 0.89)
RP concentration (mg P L^{-1}) median (10^{th}; 90^{th})	0.03 (0.04; 0.10)	0.05 (0.04; 0.11)	0.02 (0.01; 0.09)	0.01 (0.00; 0.04)
RP/TP ratio during recorded storm events (%)	40 to 60	30 to 60	10 to 80	<10
$f_{L90\%}$ (TP ; RP)	51 ; 54	21 ; 34	-	-
number of storm events per year	38	49	38	47
average event duration (h)	42	43	30	18
% of events with amplitude under 0.1 mm h^{-1}	61	51	71	88
% of events with duration under 3 days	87	94	95	97

Table 3. Percentiles 10, 50 and 90 of relative RMSE on fluxes computed for all identified storm events using non-linear model M3 after 500 simulations for total phosphorus (TP) and reactive phosphorus (RP).

	Timoleague	Ballycanew	Kervidy-Naizin	Moulinet
TP - %RMSE median (10^{th}; 90^{th})	51 (33; 76)	75 (60; 93)	79 (48; 129)	104 (53; 193)
RP - %RMSE median (10^{th}; 90^{th})	77 (48; 346)	72 (39; 177)	79 (54; 287)	238 (118; 1356)

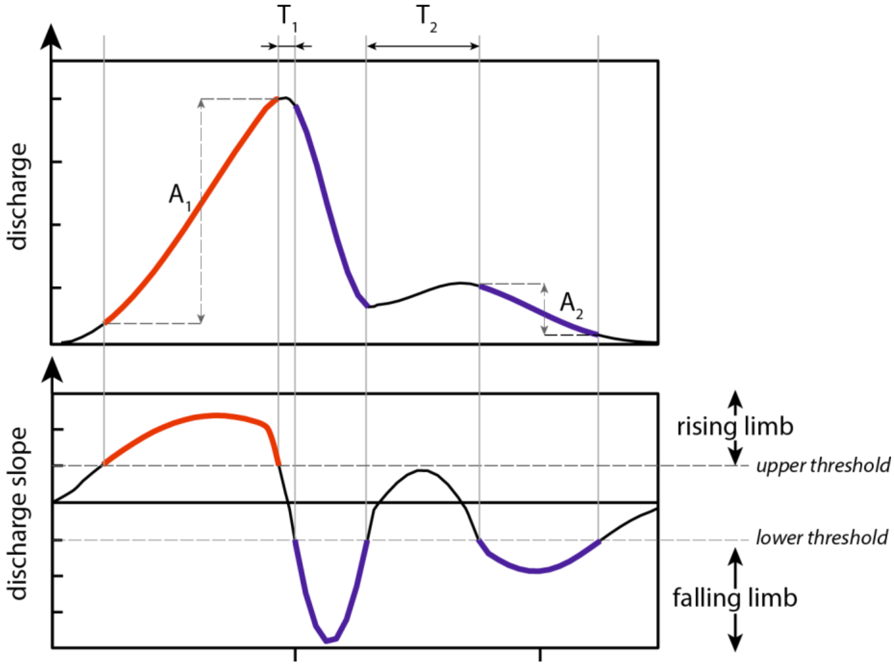
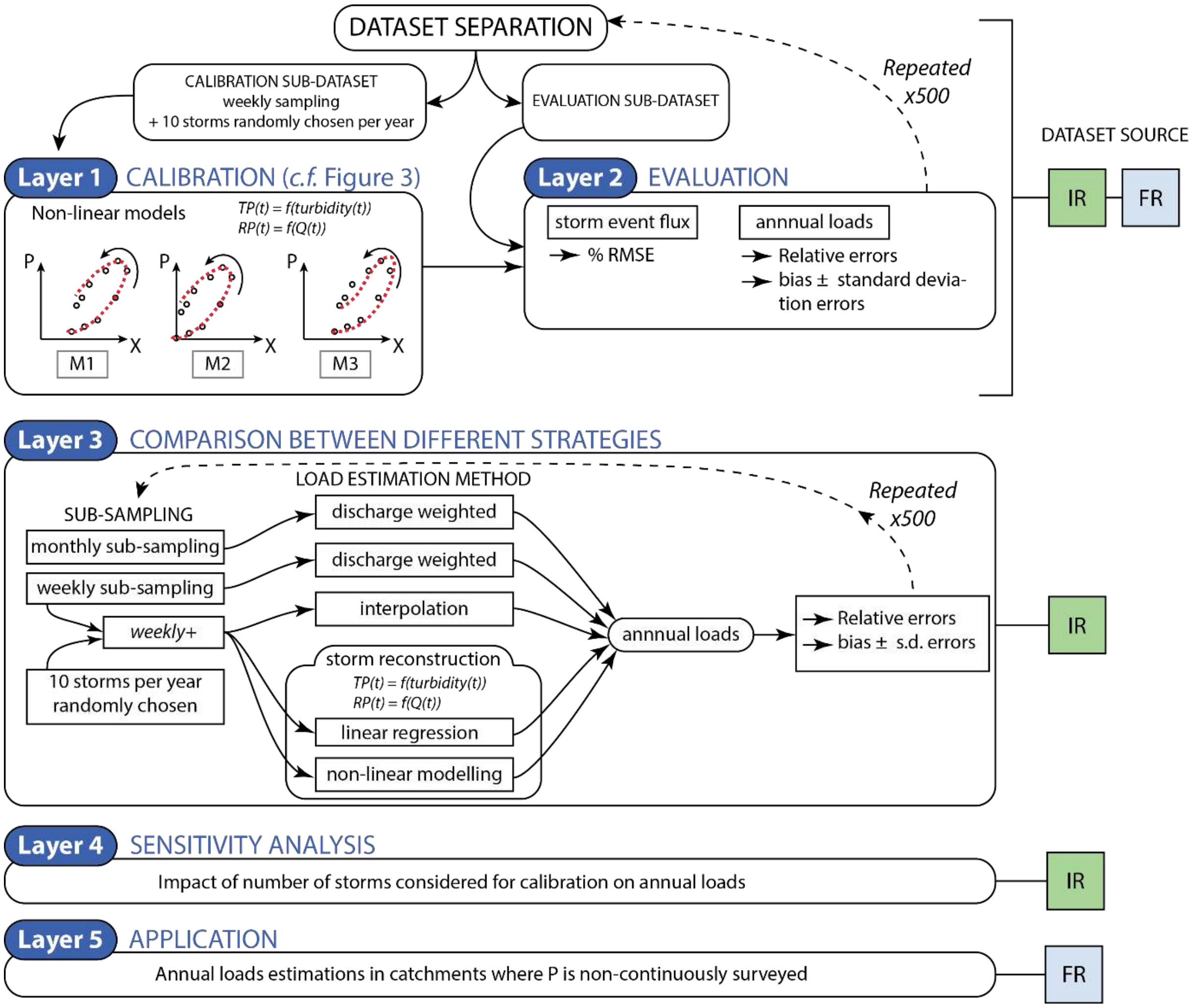


Figure 1. Conceptual view of the algorithm developed to identify a storm event in discharge time series. A_i: storm event amplitudes, T_i: time between two identified stages.



787

788 **Figure 2.** Successive layers of analysis included in this study. Capital letters on the right side indicate
789 the source of dataset used for the corresponding layer: IR corresponds to the Irish datasets; FR to the
790 French datasets.

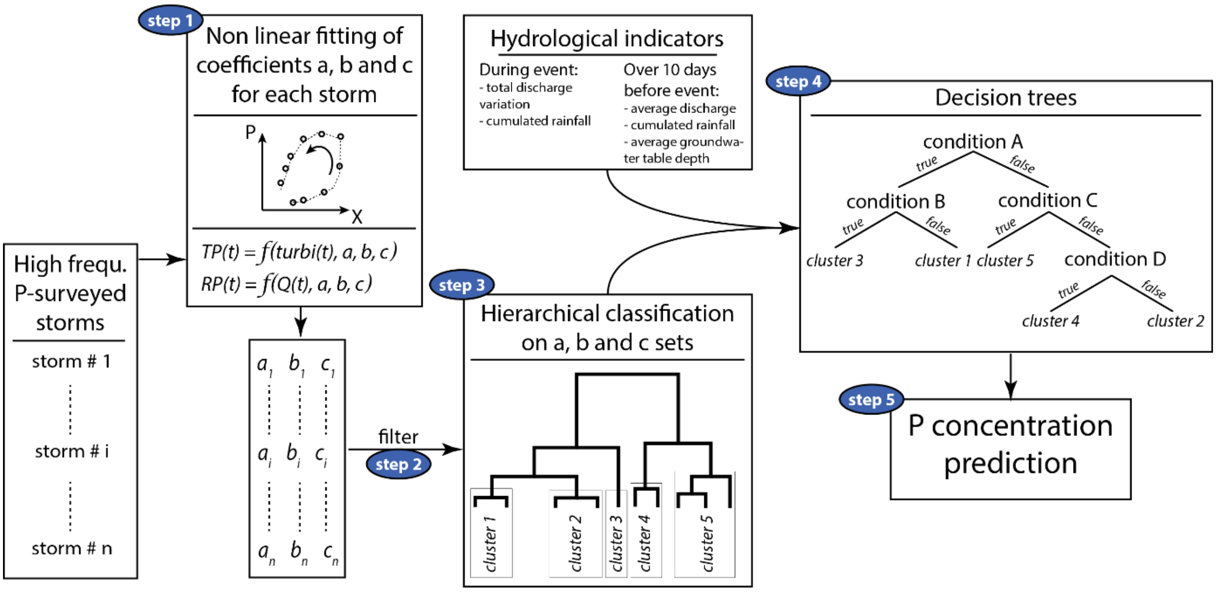
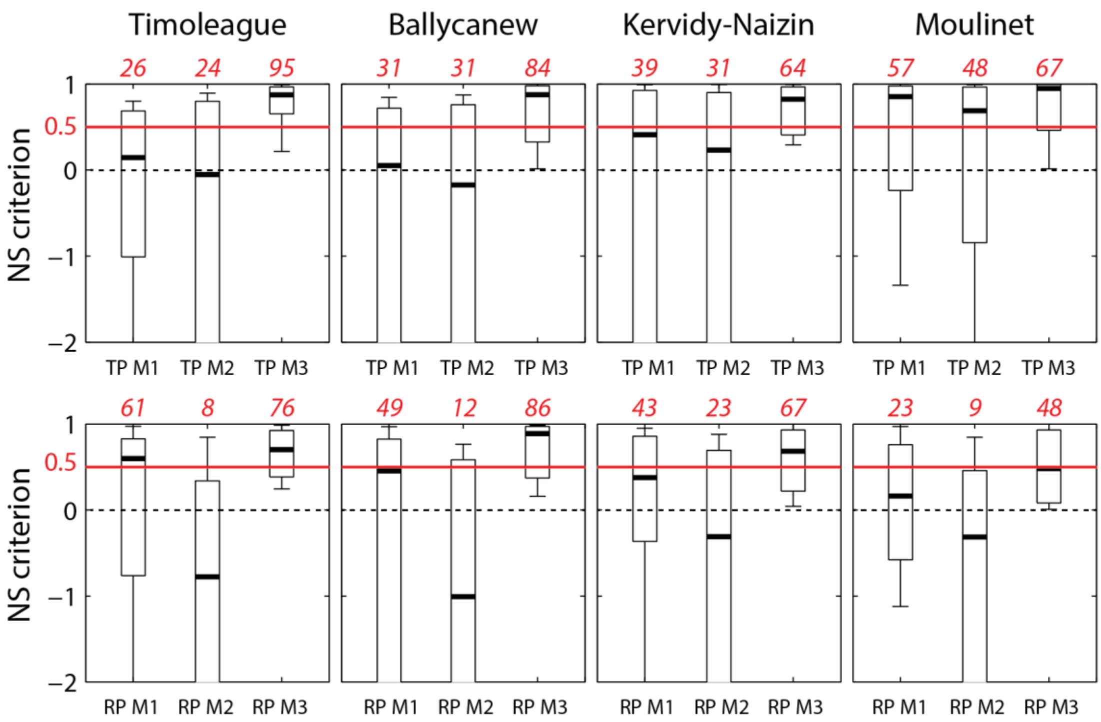


Figure 3. Successive steps for building non-linear empirical models



794 **Figure 4.** Performance during calibration step of non-linear models. Nash-Sutcliffe criterion for all P-surveyed events during calibration of non-linear empirical models M1, M2, M3. Red italic numbers represent the percentage of surveyed storms with NS criterion > 0.5.

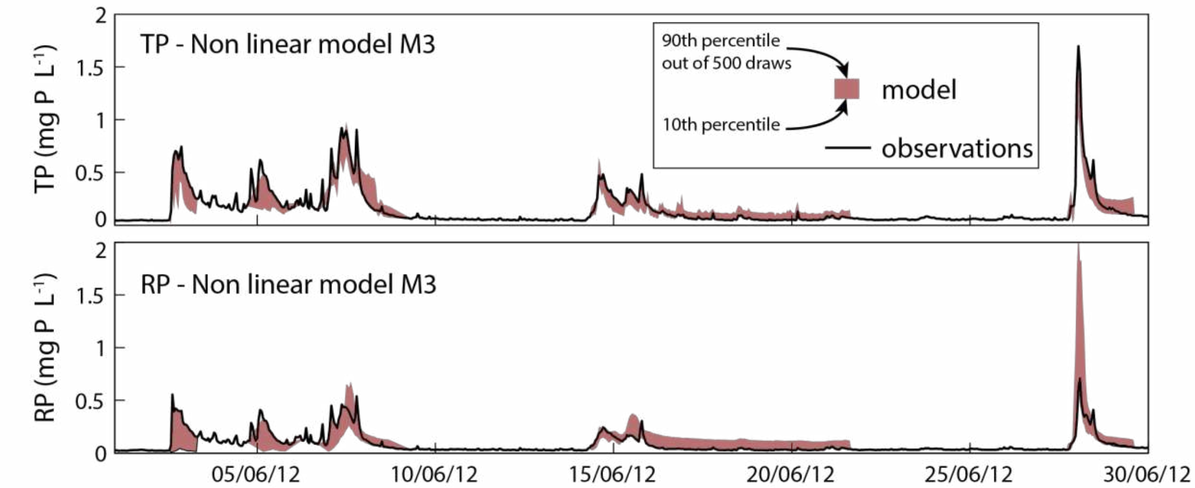


Figure 5. Example of continuous TP and RP concentration series after storm reconstruction based on the non-linear model M3, during June 2012 in the Timoleague catchment.

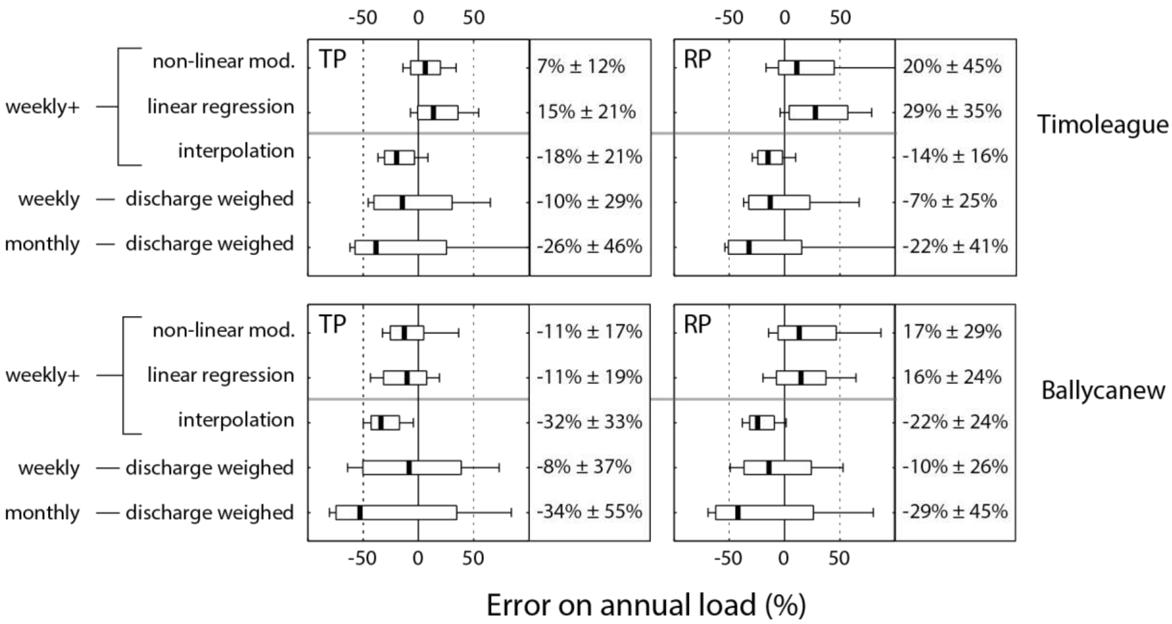


Figure 6. TP and RP relative errors on annual load estimations using non-linear modelling, a simple linear regression model, interpolation based on a *weekly+* survey, and discharge weighted method based on weekly or monthly sampling strategies. Relative bias ± s.d. errors are indicated on the right axis of each panel.

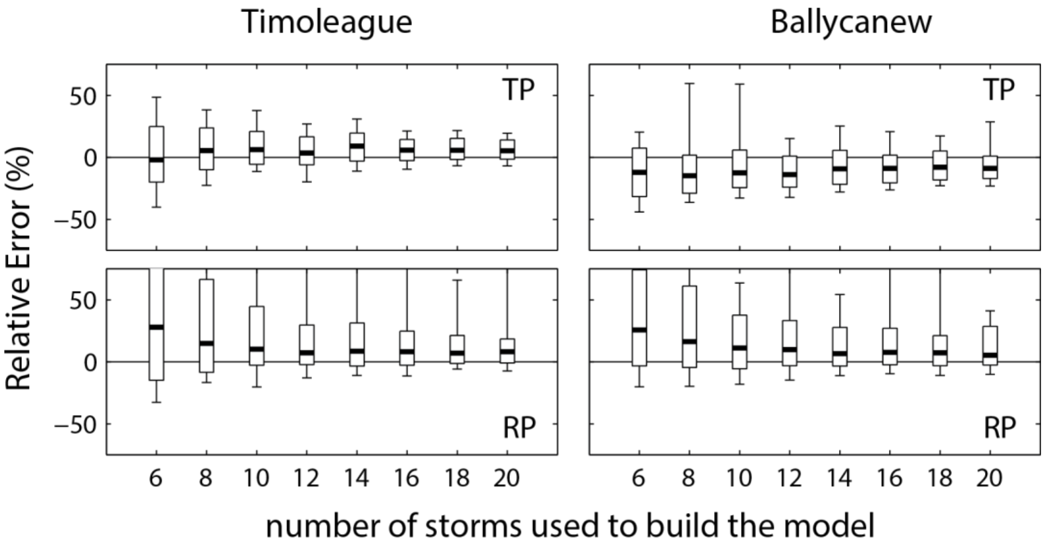


Figure 7. Sensitivity of the annual load estimations to the number of events initially used to calibrate non-linear model M3 at Timoleague and Ballycanew (500 random draws).

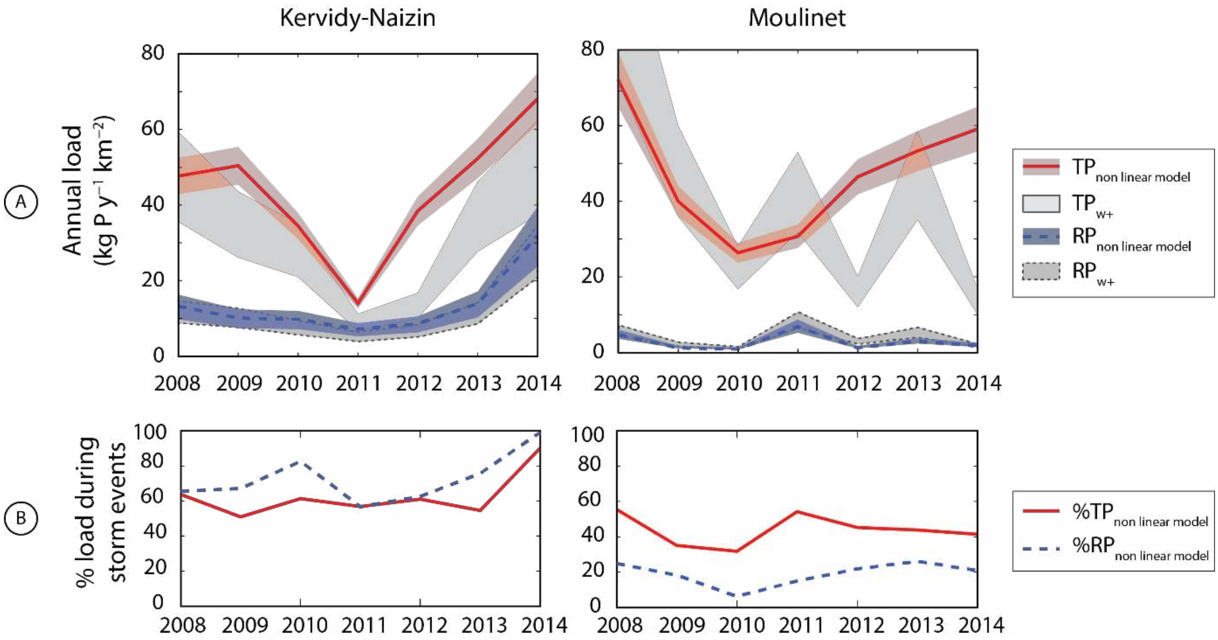


Figure 8. A) Application of the non-linear empirical method M3 to estimate annual TP and RP loads and compared to estimations based on a *weekly+* survey without storm event reconstruction in Kervidy-Naizin and Moulinet catchments. Uncertainty ranges are based on results from Irish datasets. B) Proportion of load occurring during storm events only.